Bayesian model determination in complex systems

Mohammadi, Abdolreza

Chapter 3

Bayesian Modelling of Dupuytren Disease Using Gaussian Copula Graphical Models

3.1 Abstract

Dupuytren disease is a fibroproliferative disorder with unknown etiology that often progresses and eventually can cause permanent contractures of the affected fingers. Most of the research on severity of the disease and the phenotype of this disease are observational studies without concrete statistical analyses. There is a lack of multivariate analysis for the disease taking into account potential risk factors. In this paper, we provide a novel Bayesian framework to discover potential risk factors and investigate which fingers are jointly affected. Gaussian copula graphical modelling is one potential way to discover the underlying conditional independence of variables in mixed data. Our Bayesian approach is based on Gaussian copula graphical models. We embed a graph selection procedure inside a semiparametric Gaussian copula. We carry out the posterior inference by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on a birth-death process. We implemented the method in the R package `BDgraph`.

Key words: Dupuytren disease; Risk factors; Bayesian inference; Gaussian copula graphical models; Bayesian model selection; Latent variable models; Birth-death process; Markov chain Monte Carlo.

3.2 Introduction

Dupuytren disease is inherited and presents worldwide, which is more prevalent in people with northern European ancestry (3). This disease is an incurable fibroproliferative disorder that alters the palmar hand and may causes progressive and permanent flextion contracture of the fingers. At the first stage of the disease, giving rise to the development of skin pitting and subcutaneous nodules in the palm; see Figure 3.1 in the left side. At a later stage, cords appear that connect the nodules and may contract the fingers into a flexed position; see Figure 3.1 in the right side. A contracture can arise isolated in a single ray or even multiple rays and affects them in decreasing order. The disease mostly appears in the ulnar side of the hand, i.g. it affects the little and ring fingers most-frequently, see figure 3.2. The only treatment available is surgical interventions. The main questions are: (1) Can we determine what variables affect the disease and how? (2) Do we need to intervene on multiple fingers when a surgical intervention takes place? The first question is more of an epidemiological question, while the second question is a clinical question.

Empirical research has described the patterns of occurrence of Dupuytren disease in multiple fingers. (19) have stated that most often the combination of affected ring and little fingers occurred, followed by the combination of an affected third, fourth, and fifth finger. (31) found that Dupuytren disease rarely affects isolated radial side, and that radial effect often associate with an affected ulnar side. (20) noticed that patients who had required surgery because of firmly affected thumb were on average 8 years older and had suffered significantly longer from the disease compared with patients with a mildly affected radial side. Moreover, these patients suffered from ulnar disease that repeatedly had required surgery, suggesting an intractable form of disease. More recently, (13), with a multivariate ordinal logit model, suggested that the middle finger is substantially correlated with other fingers on the ulnar side, and the thumb and index finger are correlated. They taking into account age and sex, and tested hypotheses on independence between groups of fingers. However, there is a lack of multivariate analysis study for the disease with taking into account potential risk factors.

Essential risk factors of Dupuytren disease is has been variably attributed to both phenotypic and genotypic factors (29). Essential risk factors include genetic predisposition and ethnicity, as well as sex and age. Research on family studies and twin studies recommended that Dupuytren disease has a potential genetic risk factors. However, until now, it is unclear whether Dupuytren disease is a complex oligogenic or a simple monogenic Mendelian disorder. Several life-style risk factors (some considered controversial) include

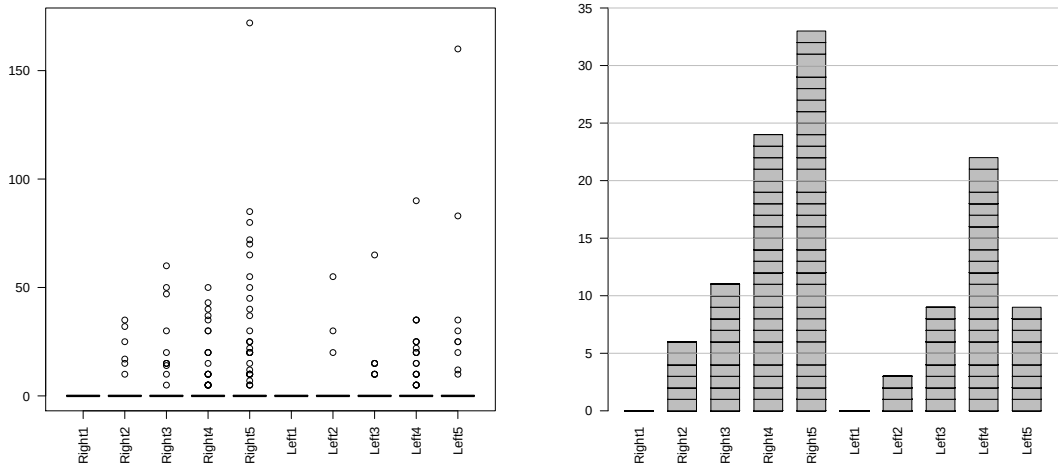
(a) With a flexed position (b) Without a flexed position

Fig. 3.1 In the left is a hand image of a patient who has Dupuytren disease and his finger has been affected by the disease. Palmar nodules and small cords without signs of contracture. In the right is a hand image of a patient who has Dupuytren disease and his fingers has not been affected by the disease yet.

smoking, excessive alcohol consumption, manual work, and hand trauma, but also several diseases, such as diabetes mellitus and epilepsy, are thought to play a role in the cause of Dupuytren disease. However, the role of these risk factors and diseases has not been fully elucidated, and the results of different studies are occasionally conflicting (12).

In this paper we analyse data which are collected in the north of Netherlands from patients who have Dupuytren disease. Both hands of patients are examined for signs of Dupuytren disease. These are tethering of the skin, nodules, cords, and finger contractures in patients with cords. Severity of the disease are measured by total angles on each of the 10 fingers. For the potential risk factors, in addition, we inquired about smoking habits, alcohol consumption, whether participants had performed manual labor during a significant part of their life, and whether they had sustained hand injury in the past, including surgery. In addition, we inquired about the presence of Ledderhose diabetes, epilepsy, peyronie, knucklepade, and liver disease; familial occurrence of Dupuytren disease, defined as a first-degree relative with Dupuytren disease. In our dataset we have 279 patients in which 79 of them, have the disease in at least one of their fingers. Therefore, there are a lot of zeros in the data as shown in figure 3.2. Beside, there are 13 potential risk factors. This mixed data set contains binary (disease factors), discrete (alcohol and hand injury), and continuous variables (total angles for fingers).

The primary aim of this paper is to model the relationships between the risk factors



(a) Boxplot (b) Histogram

Fig. 3.2 In the left is a boxplot for all hand fingers which is based on the total angles of the fingers. In the right is is occurrence of rays affected with Dupuytren disease for all 10 hand fingers.

and disease indicators for Dupuytren disease based on this mixed dataset. We propose an efficient Bayesian statistical methodology based on Gaussian copula graphical models that can be applied to binary, ordinal or continuous variables simultaneously. We embed the graphical model inside a semiparametric framework, using extended rank likelihood (10). We carry out posterior inference for the graph structure and the precision matrix by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on a continuous-time birth-death process (22).

Graphical models provide an effective way to describe statistical patterns in data. In this context undirected Gaussian graphical models are commonly used, since inference in such models is often tractable. In undirected Gaussian graphical models, the graph structure is characterized by its precision matrix (the inverse of covariance matrix): the non-zero entries in the precision matrix show the edges in the graph. In the real world, data are often non-Gaussian. For non-Gaussian continuous data, variables can be transformed to Gaussian latent variables. For discrete data, however, the situation is more convoluted; there is no one-to-one transformation into latent Gaussian variables. A common approach is to apply a Markov chain Monte Carlo method (MCMC) to simulate both the latent Gaussian variables and the posterior distributions (10). A related approach is the Gaussian copula graphical models developed by (7). In their method, they have designed the sampler algorithm which is based on reversible-jump MCMC approach. Here, in our proposed method

we implement the birth-death MCMC approach (22) which has several computational advantages compared with reversible-jump MCMC approach; see (22, Section 4).

The paper is organized as follows. In Section 3.3, we introduce a comprehensive Bayesian framework based on Gaussian copula graphical models. In addition, we show the performance of our methodology and we compare it with state-of-the-art alternatives. In section 3.4 we analyses Dupuytren disease dataset based on our proposed Bayesian statistical methodology. In this section, first we discover the potential phenotype risk factors for the Dupuytren disease. Moreover, we consider the severity of Dupuytren disease between pairs of 10 hand fingers; the result help surgeons to decide weather they should operate one finger or they should operate multiple fingers simultaneously. In the last section, we discuss the connections to existing methods and possible future directions in the last section.

3.3 Methodology

3.3.1 Gaussian graphical models

In graphical models, conditional dependence relationships among random variables are presented as a graph G. A graph $G = (V, E)$ specifies a set of vertices $V = \{1, 2, ..., p\}$, where each vertex corresponds to a random variable, and a set of existing edges $E \subset V \times V$ (15). $\overline{E}$ denotes the set of non-existing edges. We focus here on undirected graphical models in where $(i, j) \in E$ is equivalent with $(j, i) \in E$, also known as Markov random fields. The absence of an edge between two vertices specifies the pairwise conditional independence of these two variables given the remaining variables, while an edge between the two variables determines the conditional dependence of the variables.

For a p-dimensional variable exists in total $2^{p(p-1)/2}$ possible conditional independence graphs. Even with a relatively small number of variables, the size of graph space is enormous. The graph space can be explored by stochastic search algorithms (22, 8, 11). These types of algorithms explore the graph space by adding or deleting one edge at each step, known as a neighborhood search algorithm.

A graphical model that follows a multivariate normal distribution is called a Gaussian graphical models, also known as a covariance selection model (6). Zero entries in the precision matrix correspond to the absence of edges on the graph and conditional independence between pairs of random variables given all other variables. We define a zero

mean Gaussian graphical model with respect to the graph G as

$$\mathcal{M}_G = \left\{ \mathcal{N}_p(0, \Sigma) \mid K = \Sigma^{-1} \in \mathbb{P}_G \right\},$$

where $\mathbb{P}_G$ denotes the space of $p \times p$ positive definite matrices with entries (i, j) equal to zero whenever $(i, j) \in \overline{E}$. Let $\mathbf{z} = (\mathbf{z}^1, ..., \mathbf{z}^n)$ be an independent and identically distributed sample of size n from model $\mathcal{M}_G$, where $\mathbf{z}^i$ is a p dimensional vector of variables. Then, the likelihood is

$$P(\mathbf{z}|K, G) \propto |K|^{n/2} \exp\left\{ -\frac{1}{2}\text{tr}(KS) \right\}, \tag{3.1}$$

where $S = \mathbf{z}'\mathbf{z}$.

3.3.2 Gaussian copula graphical models

A copula is a multivariate cumulative distribution function whose uniform marginals are on the interval $[0, 1]$. Copulas provide a flexible tool for understanding dependence among random variables, in particular for non-Gaussian multivariate data. By Sklar's theorem (30) there exists a copula C such that any p dimensional distribution function H can be completely specified by its marginal distributions and a copula C satisfying

$$H(y_1, \ldots, y_p) = C\left(F_1(y_1), \ldots, F_p(y_p)\right),$$

where F_j are the univariate marginal distributions of H. If all F_j are all continuous, then C is unique, otherwise it is uniquely determined on $\text{Ran}(F_1) \times \cdots \times \text{Ran}(F_p)$ which is the cartesian product of the ranges of F_j. Conversely, a copula function can be extracted from any p dimension distribution function H and marginal distributions F_j by

$$C(u_1, \ldots, u_p) = H\left(F_1^{-1}(y_1), \ldots, F_p^{-1}(y_p)\right),$$

where $F_j^{-1}(s) = \inf\{t \mid F_j(t) \geq s\}$ are the pseudo-inverse of F_j.

The decomposition of a joint distribution into marginal distributions and a copula suggests that the copula captures the essential features of dependence between random variables. Moreover, the copula measure of dependence is invariant to any monotone transformation of random variables. Thus, copulas allow one to model the marginal distributions and the dependence structure of a multivariate random variables separately. In copula modelling, Genest et al. (9) develop a popular semiparametric estimation approach

or rank likelihood based estimation in which the association among variables is represented with a parametric copula but the marginals are treated as nuisance parameters. The marginals are estimated non-parametrically using the scaled empirical distribution function $\hat{F}_j(y) = \frac{n}{n+1}F_{n_j}(y)$, where $F_{n_j}(y) = \frac{1}{n}\sum_{i=1}^{n} I\{y_{ij} \leq y\}$. As a result estimation and inference are robust to misspecification of marginal distributions.

The semiparametric estimators are well-behaved for continuous data but fail for discrete data, for which the distribution of the ranks depends on the univariate marginal distributions, making them somewhat inappropriate for the analysis of mixed continuous and discrete data (10). To remedy this, Hoff (10) propose the extended rank likelihood which is a type of marginal likelihood that does not depend on the marginal distributions of the observed variables. Under the extended rank likelihood approach the ranks are free of the nuisance parameters (or marginal distributions) of the discrete data. This makes the extended rank likelihood approach more focused on the determination of graphical models (or multivariate association) and avoids the difficult problem of modelling the marginal distributions (7).

In case of ordered discrete and continuous variables, a Gaussian copula has been considered to describe dependence pattern between heterogeneous variables using the extended rank likelihood in Gaussian copula graphical modelling (7). Let Y be a collection of continuous, binary, ordinal or count variables with F_j the marginal distribution of Y_j and F_j^{-1} its pseudo inverse. For constructing a joint distribution of Y, we introduce a multivariate normal latent variable as follows

$$Z_1, ..., Z_n \overset{iid}{\sim} \mathcal{N}(0, \Gamma),$$

and define the observed data as

$$Y_{ij} = F_j^{-1}(\Phi(Z_{ij})),$$

where Γ is the correlation matrix for a Gaussian copula. The joint distribution of $\mathbf{Y}$ is given by

$$P\left(Y_1 \leq y_1, \ldots, Y_p \leq y_p\right) = C(F_1(y_1), \ldots, F_p(y_p) \mid \Gamma),$$

where $C(\cdot)$ is the Gaussian copula given by

$$C(u_1, \ldots, u_p \mid \Gamma) = \Phi_p\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p) \mid \Gamma\right),$$

where $\Phi_p(\cdot)$ is the cumulative distribution of a multivariate normal distribution and $\Phi(\cdot)$ is a cumulative distribution function of a univariate normal distribution. Hence the joint cumulative distribution function is

$$P\big(Y_1 \leq y_1, \ldots, Y_p \leq y_p\big) = \Phi_p\big(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F(y_p)) \,|\, \Gamma\big). \tag{3.2}$$

In the semiparametric copula estimation, since the marginals are treated as nuisance parameters the joint distribution in (4.9) is parametrized only by the correlation matrix of the Gaussian copula, Γ.

Our aim is to infer the underlying graph structure G of the observed variables $\mathbf{Y}$ implied by the continuous latent variables $\mathbf{Z}$. Since $\mathbf{Z}$s are unobservable we follow the idea of (10) that relate them to the observed data as follows. Given the observed data Y from a sample of n observations, the latent samples $\mathbf{z}$ are constrained to belong to the set

$$\mathcal{A}(\mathbf{y}) = \{\mathbf{z} \in \mathbb{R}^{n \times p} : l_j^r(\mathbf{z}) < z_j^{(r)} < u_j^r(\mathbf{z}), r = 1, \ldots, n; j = 1, \ldots, p\}.$$

where

$$\begin{aligned} l_j^r(\mathbf{z}) &= \max\left\{z_j^{(k)} : y_j^{(s)} < y_j^{(r)}\right\}, \\ u_j^r(\mathbf{z}) &= \min\left\{z_j^{(s)} : y_j^{(r)} < y_j^{(s)}\right\}. \end{aligned} \tag{3.3}$$

Further (10) suggests that inference on the latent space can be performed by substituting the observed data $\mathbf{y}$ with the event $\mathcal{A}(\mathbf{y})$. For a given graph G and precision matrix $K = \Gamma^{-1}$, the likelihood is defined as:

$$\begin{aligned} P(\mathbf{y} \mid K, G, F_1, ..., F_p) &= P(\mathbf{y}, \mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid K, G, F_1, ..., F_p) \\ &= P(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid K, G) \\ &\quad \times P(\mathbf{y} \mid \mathbf{z} \in \mathcal{A}(\mathbf{y}), K, G, F_1, ..., F_p). \end{aligned}$$

The only part of the observed data likelihood relevant for inference on K and G is $P(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid K, G)$. Thus, the extended rank likelihood function as referred by (10) is given by

$$P(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \,|\, K, G) = P(\mathbf{z} \in \mathcal{A}(\mathbf{y}) \mid K, G) = \int_{\mathcal{A}(\mathbf{y})} P(\mathbf{z} \,|\, K, G) d\mathbf{z},$$

where the expression inside the integral for the Gaussian copula based distribution given by (4.9) takes a similar form as in 5.2.

Therefore, we can infer about (K, G) by obtaining a posterior distribution

$$P(K, G|\mathbf{z} \in \mathcal{A}(\mathbf{y})) \propto P(\mathbf{z} \in \mathcal{A}(\mathbf{y})|K, G)P(K \mid G)P(G),$$

which is discussed in detail in the next sections. Moreover, we evaluate the results induced by the latent variables using posterior predictive analysis on the scale of the original mixed variables.

3.3.3 Bayesian Gaussian copula graphical models

Prior specification

In this section we discuss the specification of prior distributions for the graph G and the precision matrix K. For the prior distribution of the graph, we propose to use the discrete uniform distribution over the graph space, $P(G) \propto 1$, as a non-informative prior. Other choices of priors for the graph structure have been considered by modelling the joint state of the edges (28), encouraging sparse graphs (11) or a truncated Poisson distribution on the graph size (22).

We consider the G-Wishart (27) distribution for the prior distribution of the precision matrix. The G-Wishart distributions is conjugate for normally distributed data and places no probability mass on zero entries of the precision matrix. Matrix $K \in \mathbb{P}_G$ has the G-Wishart distribution $W_G(b, D)$, if

$$P(K|G) = \frac{1}{I_G(b, D)}|K|^{(b-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(DK)\right\},$$

where $b > 2$ is the degree of freedom, D is a symmetric positive definite matrix, and $I_G(b, D)$ is the normalizing constant,

$$I_G(b, D) = \int_{\mathbb{P}_G} |K|^{(b-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(DK)\right\} dK.$$

If graph G is complete the G-Wishart distribution reduces to the usual Wishart distribution. In that case, its normalizing constant has an explicit form (24). If a graph is decomposable, $I_G(b, D)$ can be calculated explicitly (27). For non-decomposable graphs, we can approximate $I_G(b, D)$ by a Monte Carlo approach (2) or a Laplace approximation (17).

The G-Wishart prior is conjugate to the likelihood (5.2), hence, the posterior distribu-

tion of K is

$$P(K|Z \in \mathcal{A}(\mathbf{y}), G) = \frac{1}{I_G(b^*, D^*)}|K|^{(b^*-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(D^*K)\right\},$$

where $b^* = b + n$ and $D^* = D + S$, that is, $W_G(b^*, D^*)$. For other choices of priors for the precision matrix see (36, 34, 33, 37).

Posterior inference

Consider the joint posterior distribution of $K \in P_G$ and the graph G given by

$$P(K, G \mid Z \in \mathcal{A}(\mathbf{y})) \propto P(Z \in \mathcal{A}(\mathbf{y}) \mid K)\, P(K \mid G)\, P(G). \tag{3.4}$$

Sampling from this joint posterior distribution can be done by a computationally efficient birth-death MCMC sampler proposed in Mohammadi and Wit (2013) for Gaussian graphical models. Here we extend their algorithm for the more general case of Gaussian copula graphical models. Our algorithm is based on a continuous time birth-death Markov process in which the algorithm explores the graph space by adding or removing an edge in a birth or death event. The birth and death rates of edges occur in continuous time with the rates determined by the stationary distribution of the process. The algorithm is considered in such a way that the stationary distribution equals the target joint posterior distribution of the graph and the precision matrix (5.3).

The birth-death process is designed in such a way that the birth and death events are independent Poisson processes; the time between two successive events has an exponential distribution. Therefore, the probability of birth and death events are proportional to their rates.

Mohammadi and Wit (22, section 3) prove that by considering the following birth and death rates, the birth-death MCMC sampling algorithm converges to the target joint posterior distribution of the graph and the precision matrix,

$$\beta_e(K) = \frac{P(G^{+e}, K^{+e} \setminus (k_{ij}, k_{jj}) | Z \in \mathcal{A}(\mathbf{y}))}{P(G, K \setminus k_{jj} | Z \in \mathcal{A}(\mathbf{y}))}, \quad \text{for each } \; e \in \overline{E}, \tag{3.5}$$

$$\delta_e(K) = \frac{P(G^{-e}, K^{-e} \setminus k_{jj} | Z \in \mathcal{A}(\mathbf{y}))}{P(G, K \setminus (k_{ij}, k_{jj}) | Z \in \mathcal{A}(\mathbf{y}))}, \qquad \text{for each } \; e \in E, \tag{3.6}$$

in which $G^{+e} = (V, E \cup \{e\})$, and $K^{+e} \in \mathbb{P}_{G^{+e}}$ and similarly $G^{-e} = (V, E \setminus \{e\})$, and $K^{-e} \in \mathbb{P}_{G^{-e}}$. The extended birth-death MCMC algorithm for Gaussian copula graphical

models are summarized in Algorithm 3.1.

1 **Algorithm 3.1** Given a graph $G = (V, E)$ with a precision matrix K, iterate the following steps:

1. Sample the latent data. For each $r \in V$ and $j \in \{1, 2, ..., n\}$, we update the latent value $z_r^{(j)}$ from its full conditional distribution

$$Z_r|K, Z_{V\setminus\{r\}} = z^{(j)}_{K,V\setminus\{r\}} \sim \mathcal{N}\left(-\sum_{r'} K_{rr'} z_{r'}^{(j)}/K_{rr}, 1/K_{rr}\right),$$

truncated to the interval $\left[L_r^j, U_r^j\right]$ in (4.10).

2. Sample the graph based on birth and death process.

2.1. Calculate the birth rates by equation 5.7 and $\beta(K) = \sum_{e\in\overline{E}} \beta_e(K)$,

2.2. Calculate the death rates by equation 5.8 and $\delta(K) = \sum_{e\in E} \delta_e(K)$,

2.3. Calculate the waiting time by $\mathcal{W}(K) = 1/(\beta(K) + \delta(K))$,

2.4. Calculate the type of jump (birth or death),

3. Sample the new precision matrix, according to the type of jump, based on Algorithm 3.2.

In Algorithm 3.1, the first step is to sample from the latent variables given the observed data. Then, based on this sample, we calculate the birth and death rates and waiting times. Based on birth and death rates we calculate the type of jump. Details of how to efficiently calculate the birth and death rates are discussed in subsection 1. Finally in step 3, according to the new state of jump, we sample from new precision matrix using a direct sampling scheme from the G-Wishart distribution which is described in Algorithm 3.2. in subsection 1.

To calculate the posterior probability of a graph we compute the Rao-Blackwellized sample mean (4, subsection 2.5). The Rao-Blackwellized estimate of the posterior graph probability is the proportion of the total waiting time for that graph (see Figure 4.2 in the right). The weights are equal to the length of the waiting time in each state (e.g. $\{W_1, W_2, W_3, ...\}$ in Figure 4.2).

Computing the birth and death rates

Calculating the birth and death rates (5.7 and 5.8) is the bottleneck of our BDMCMC algorithm. Here, we explain how to calculate efficiently the death rates; the birth rates are calculated a similar manner.

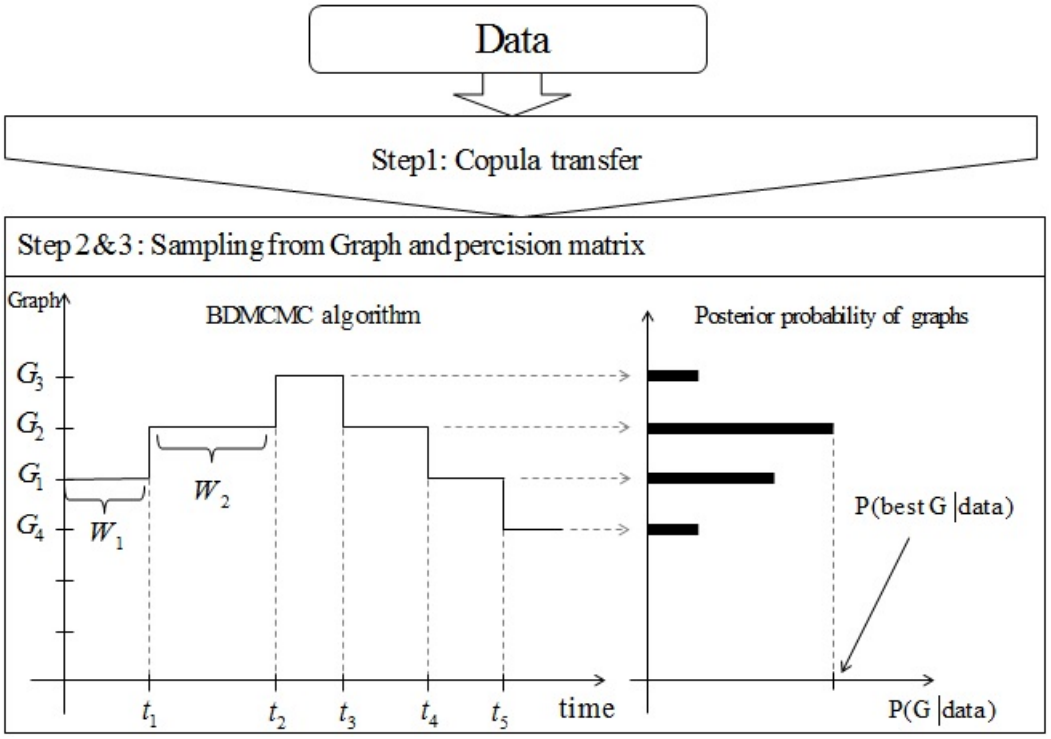


Fig. 3.3 This image visualizes Algorithm 3.1. (Bottom left) Continuous time BDMCMC algorithm where $\{W_1, W_2, ...\}$ denote waiting times and $\{t_1, t_2, ...\}$ denote jumping times. (Bottom right) Estimated posterior probability of the graphs which are proportional to sum of their waiting times.

Following (22) and after some simplification, for each $e = (i, j) \in E$, we have

$$\delta_e(K) = \frac{P(G^{-e})}{P(G)} \frac{I_G(b, D)}{I_{G^{-e}}(b, D)} \left(\frac{D^*_{jj}}{2\pi(k_{ii} - k^1_{11})}\right)^{\frac{1}{2}} H(K, D^*), \tag{3.7}$$

where

$$H(K, D^*) = \exp\left\{-\frac{1}{2}\left[\text{tr}(D^*_{e,e}(K^0 - K^1)) - (D^*_{ii} - \frac{(D^*_{ij})^2}{D^*_{jj}})(k_{ii} - k^1_{11})\right]\right\},$$

in which

$$K^0 = \begin{bmatrix} k_{ii} & 0 \\ 0 & K_{j,V\setminus j}(K_{V\setminus j, V\setminus j})^{-1} K_{V\setminus j, j} \end{bmatrix},$$

and $K^1 = K_{e,V\setminus e}(K_{V\setminus e, V\setminus e})^{-1} K_{V\setminus e, e}$. The computational bottleneck in (3.7) is the ratio of normalizing constants.

Dealing with calculation of normalizing constants Calculating the ratio of normalizing constants has been a major issue in recent literature (32, 35, 22). To compute the normalizing constants of a G-Wishart, (27) proposed an importance sampling algorithm, while (2) developed a Monte Carlo method. These methods can be computationally expensive and numerical instable (11, 35). (35, 5, 22) developed an alternative approach, which borrows ideas from the exchange algorithm (25) and the double Metropolis-Hastings algorithm (18) to compute the ratio of such normalizing constants. When the dimension of the problem is high, the curse of dimensionality may be a serious difficulty for the double MH sampler (18). More recently (32) derived an explicit representation of the normalizing constant ratio.

Theorem 3.1 (Uhler et al. 32). *Let $G = (V, E)$ be an undirected graph and $G^{-e} = (V, E^{-e})$ denotes the graph G with one less edge e. Then*

$$\frac{I_G(b, \mathbb{I}_p)}{I_{G^{-e}}(b, \mathbb{I}_p)} = 2\sqrt{\pi}\frac{\Gamma((b+d+1)/2)}{\Gamma((b+d)/2)},$$

where d denotes the number of triangles formed by the edge e and two other edges in G and $\mathbb{I}_p$ denotes an identity matrix with p dimension.

Proof. it is immediate by using Uhler et al. (32, theorem 3.7).

Therefore, for the case of $D = \mathbb{I}_p$, we have a simplified expression for the death rates, given by

$$\delta_e(K) = \frac{P(G^{-e})}{P(G)}\frac{\Gamma((b+d+1)/2)}{\Gamma((b+d)/2)}(\frac{2D^*_{jj}}{(k_{ii} - k^1_{ii})})^{\frac{1}{2}}H(K, D^*),$$

Sampling from posterior distribution of precision matrix

Several sampling methods from a G-Wishart have been proposed; for a review of existing methods see (35) and (16). Here we use an exact sampler algorithm developed by (16) and summarized in Algorithm 3.2.

Simulation study

We perform a comprehensive simulation study with respect to different graph structures to evaluate the performance of our method and compare it to an alternative approach proposed by Dobra and Lenkoski (7), referred to as DL. We generate mixed data from a latent Gaussian copula model with 5 different types of variables, for “Gaussian”, “non-Gaussian”,

Algorithm 3.2. Direct sampler from precision matrix (16). Given a graph $G = (V, E)$ with precision matrix K:

1. Set $\Sigma = K^{-1}$,

2. Repeat for $j = 1, ..., p$, until converge:

2.1 Let $N_j \subset V$ be the set of variables that connected to j in G. Form Σ_{N_j} and $K^{-1}_{N_j,j}$ and solve

$$\hat{\beta}^*_j = \Sigma^{-1}_{N_j} K^{-1}_{N_j,j},$$

2.2 Form $\hat{\beta}_j \in R^{p-1}$ by plugging zeroes in those locations not connected to j in G and padding the elements of $\hat{\beta}^*_j$ to the rest locations,

2.3 Replace $\Sigma_{j,-j}$ and $\Sigma_{-j,j}$ with $\Sigma_{-j,-j}\hat{\beta}_j$,

3. Return $K = \Sigma^{-1}$.

"ordinal", "count", and "binary". We performed all computations with our extended R package BDgraph (21, 23).

Corresponding to different sparsity patterns, we consider 4 different kinds of synthetic graphical models:

1. *Random Graph:* A graph in which the edge set E is randomly generated from independent Bernoulli distributions with probability $2/(p-1)$ and corresponded precision matrix is generated from $K \sim W_G(3, I_p)$.
2. *Cluster Graph:* A graph in which the number of clusters is $max\left\{2, \left[p/20\right]\right\}$. Each cluster has the same structure as a random graph. The corresponded precision matrix is generated from $K \sim W_G(3, I_p)$.
3. *Scale-free Graph:* A scale-free graph has a power-low degree distribution generated by the Barabasi-Albert algorithm (1). The corresponded precision matrix is generated from $K \sim W_G(3, I_p)$.
4. *Hub Graph:* A graph in which every node is connected to one node, and corresponded precision matrix is generated from $K \sim W_G(3, I_p)$.

For each graphical model, we consider four different scenarios: (1) dimension $p = 10$ and sample size $n = 30$, (2) $p = 10$ and $n = 100$, (3) $p = 30$ and $n = 100$, (4) $p = 30$ and $n = 500$.

For each mixed data set, we fit our method and DL approach using a uniform prior for the graph and the G-Wishart prior $W_G(3, I_p)$ for the precision matrix. We run the two algorithms with the same starting points with $100,000$ iteration and $50,000$ as a burn in. Computations for this example were performed in parallel on a 235 batch nodes with 12 cores and 24 GB of memory, running Linux.

To assess the performance of the graph structure, we compute the F_1-score measure (26) for MAP graph which defined as

$$F_1\text{-score} = \frac{2\text{TP}}{2\text{TP + FP + FN}}, \tag{3.8}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. The F_1-score score lies between 0 and 1, where 1 stands for perfect identification and 0 for bad identification. Also, the mean square error (MSE) is used, defined as

$$\text{MSE} = \sum_{e} (\hat{p}_e - I(e \in G_{true}))^2, \tag{3.9}$$

where $\hat{p}_e$ is the posterior pairwise edge inclusion probabilities and $I(e \in G_{true})$ is an indicator function, such that $I(e \in G_{true}) = 1$ if $e \in G_{true}$ and zero otherwise. For our BDMCMC algorithm we calculate the posterior pairwise edge inclusion probabilities based on the Rao-Blackwellization (4, subsection 2.5) for each possible edge $e = (i, j)$ as

$$\hat{p}_e = \frac{\sum_{t=1}^{N} I(e \in G^{(t)}) \mathcal{W}(K^{(t)})}{\sum_{t=1}^{N} \mathcal{W}(K^{(t)})}, \tag{3.10}$$

where N is the number of iterations and $\mathcal{W}(K^{(t)})$ is the waiting time in the graph $G^{(t)}$ with the precision matrix $K^{(t)}$; see (22). Table 3.1 reports comparisons of our method with DL (7), where we repeat the experiments 50 times and report the average F_1-score and MSE with their standard errors in parentheses. Our method performs well overall as its F_1-score and its MSE are beter in most of the cases and mainly because of its fast convergence rate. As we expected, the DL approach converges slower compared to our method. From a theoretical point of view, both algorithms converge to the true posterior distribution, if we run them a sufficient amount of time. Thus, the results from this table just indicate how quickly the algorithms converge.

3.4 Analysis of Dupuytren disease dataset

The data set we analyses here are collected from patients who have Dupuytren disease from north of Netherlands. Both hands of patients who were willing to participate and signed an informed consent from are examined for signs of Dupuytren disease and knuckle pads. Signs of Dupuytren disease include tethering of the skin, nodules, cords, and finger contractures in patients with cords. Participants who had at least one of these features

			F1-score		MSE	
p	n	graph	BDMCMC	DL	BDMCMC	DL
10	30	Random	**0.37** (0.17)	0.33 (0.11)	**7.0** (2.0)	9.2 (1.1)
		Cluster	**0.35** (0.16)	0.30 (0.11)	**6.9** (1.8)	9.6 (1.4)
		Scale-free	0.31 (0.13)	**0.34** (0.08)	**7.6** (1.2)	9.7 (1.0)
		Hub	0.26 (0.11)	**0.31** (0.10)	**8.3** (0.8)	10.2 (0.8)
10	100	random	**0.33** (0.17)	0.32 (0.11)	**7.7** (1.7)	9.9 (1.0)
		Cluster	**0.30** (0.17)	0.28 (0.09)	**6.8** (1.5)	9.5 (1.3)
		Scale-free	**0.33** (0.16)	0.32 (0.12)	**7.3** (1.4)	9.6 (1.0)
		Hub	0.26 (0.12)	**0.31** (0.09)	**8.3** (0.9)	10.0 (1.0)
30	100	Random	**0.54** (0.06)	0.44 (0.04)	**52.3** (9.9)	59.1 (8.7)
		Cluster	**0.56** (0.05)	0.47 (0.04)	**48.0** (6.5)	54.4 (8.1)
		Scale-free	**0.53** (0.17)	0.30 (0.05)	27.7 (14.6)	**25.8** (1.7)
		Hub	**0.39** (0.08)	0.31 (0.04)	34.8 (6.8)	**30.5** (4.5)
30	500	Random	**0.79** (0.04)	0.63 (0.07)	**25.8** (6.5)	41.1 (14.3)
		Cluster	**0.79** (0.05)	0.66 (0.05)	**26.3** (5.2)	35.1 (7.9)
		Scale-free	**0.81** (0.07)	0.59 (0.06)	**9.4** (3.2)	11.7 (3.0)
		Hub	**0.73** (0.07)	0.53 (0.08)	**12.7** (3.4)	13.5 (3.2)

Table 3.1 Summary of performance measures in simulation example 1 for our method and DL (7). The table presents the F_1-score, defined in (4.13) and MSE, defined in (3.9), with 100 replications and standard deviations in parenthesis. The F_1-score reaches its best score at 1 and its worst at 0. The MSE is positive value for which 0 is minimal and smaller is better. The best models for both F_1-score and MSE are boldfaced.

were labeled as having Dupuytren disease. Severity of the disease are measured by total angles of each 10 fingers. The total angles is the sum of angles for metaccarpophalangeal joint, two interphalangeal joints (for thumb fingers are only two interphalangeal joints).

As potential risk factors, in addition, information is available about smoking habits, alcohol consumption, whether participants performed manual labor during a significant part of their life, and whether they had sustained hand injury in the past, including surgery. In addition, disease history information about the presence of Ledderhose diabetes, epilepsy, peyronie, knucklepade, or liver disease, and familial occurrence of Dupuytren disease (defined as a first-degree relative with Dupuytren disease) is available.

The data consist of 279 patients who have Dupuytren disease ($n = 279$); among those patients, 79 of them have an irreversible flexion contracture at least one of their fingers. The severity of the disease in the all 10 fingers of the patients is measured by total angles

of each fingers. To study the potential phenotype risk factors of this disease, we consider the above mentioned 14 factors.

(13) analyzes the Dupuytren disease with a multivariate ordinal logit model, taking into account age and sex, and tested hypotheses of the independence between groups of fingers. However, most of the studies on the phenotype of this disease have been observational studies without comprehensive statistical analyses.

Phenotype risk factors previously described include alcohol consumption, smoking, manual labor, hand trauma, diabetes mellitus, and epilepsy (29, 14).

In subsection 3.4.1, we infer the Dupuytren disease network with 14 potential risk factors based on our Bayesian approach. In subsection 3.4.2, we consider only the 10 fingers to infer the interaction between the fingers.

3.4.1 Inference for Dupuytren disease with risk factors

We consider the severity of disease in all 10 fingers of the patients and 14 potential phenotype risk factors of the disease, so $p = 24$. The factors are: age, sex, smoking, amount of alcohol (Alcohol), relative (Relative), number of hand injury of patients (HandInjury), Manual labour (Labour), Ledderhose disease (Ledderhose), diabetes disease (Diabetes), epilepsy disease (Epilepsy), liver Disease (LiverDisease), peyronie disease (Peyronie), knucklepade disease (Knucklepade). For each finger we measure angles of metaccarpophalangeal joint, two interphalangeal joints (for thumb fingers we only measure two interphalangeal joints); then we sum those angles for each fingers. The total angles could vary from 0 to 270 degrees; in this dataset the minimum degree is 0 and maximum 157 degrees. The age of participants (in years) ranges from 40 to 89 years, with an average age of 66 years. Smoking is binned into 3 ordered categories (never, stopped, and smoking). Amount of alcohol consumption is binned into 8 ordered categories (ranging from no alcohol to ≥ 20 consumption per week). The other variables are binary.

We apply our Bayesian framework to infer the conditional (in)dependence among the variables in order to identify the potential risk factors of the Dupuytren disease and discover how they affect the disease. We place a uniform distribution as an uninformative prior on the graph and the G-Wishart $W_G(3, I_{24})$ on the precision matrix. We run our BDMCMC algorithm for $2,000$K iterations with a $1,000$K sweeps burn-in.

The graph with the highest posterior probability is the graph with 42 edges. Figure 3.4 shows the selected graph with 26 edges, for which the posterior inclusion probabilities in (4.12) is greater then 0.5. Edges in the graph show the interactions among the genes. Figure 3.5 shows the image of all posterior inclusion probabilities for visualization.

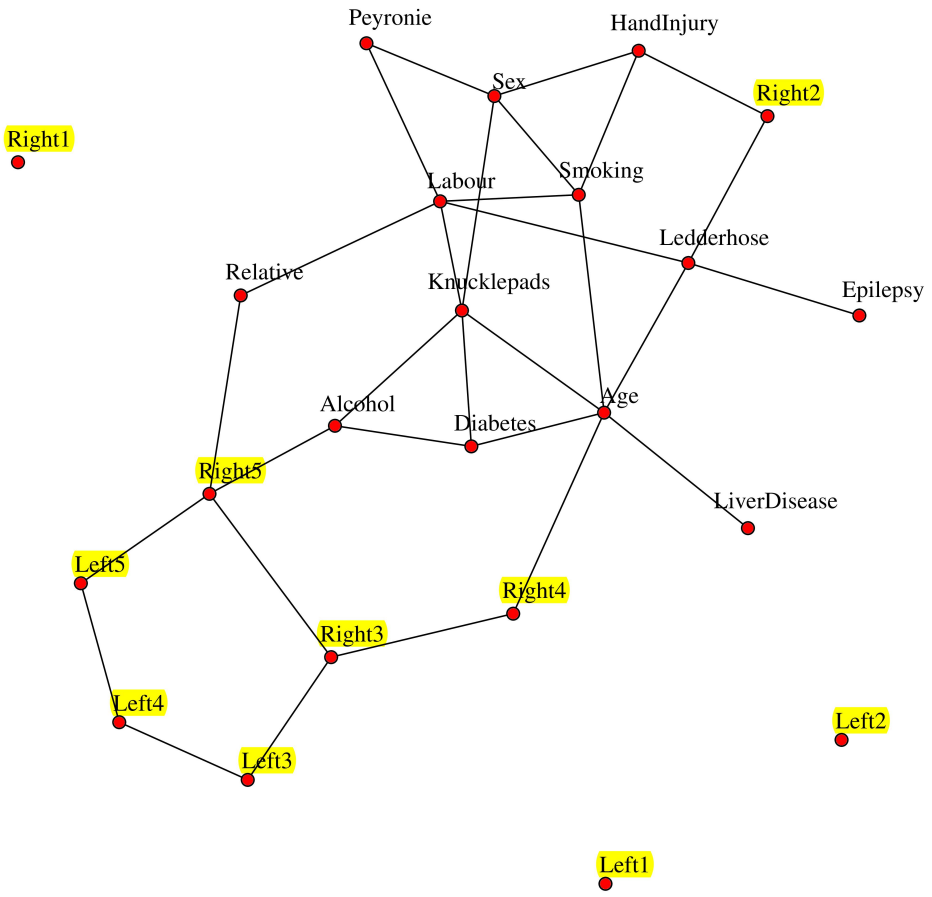


Fig. 3.4 The inferred graph for the Dupuytren disease dataset based on 14 risk factors and the total degrees of flexion in all 10 fingers. It reports the selected graph with 24 significant edges for which their posterior inclusion probabilities (4.12) are more than 0.5.

The results (Figures 3.4 and 3.5) show that factors “Age”, “Ledderhose disease”, “Hand Injury”, “Alcohol”, have a significant effect on the severity of the Dupuytren disease. Graph 3.4 also shows that factor “Age” is a hub in this graph and it plays a significant role as it affects the severity of the disease directly and indirectly through the influence of other risk factors such as “Ledderhose”.

3.4.2 Severity of Dupuytren disease between pairs of fingers

Here, we consider the severity of Dupuytren disease between pairs of 10 hand fingers. Interaction between fingers is important because it help surgeons to decide weather they should operate one finger or multiple fingers simultaneously. The main idea is that if

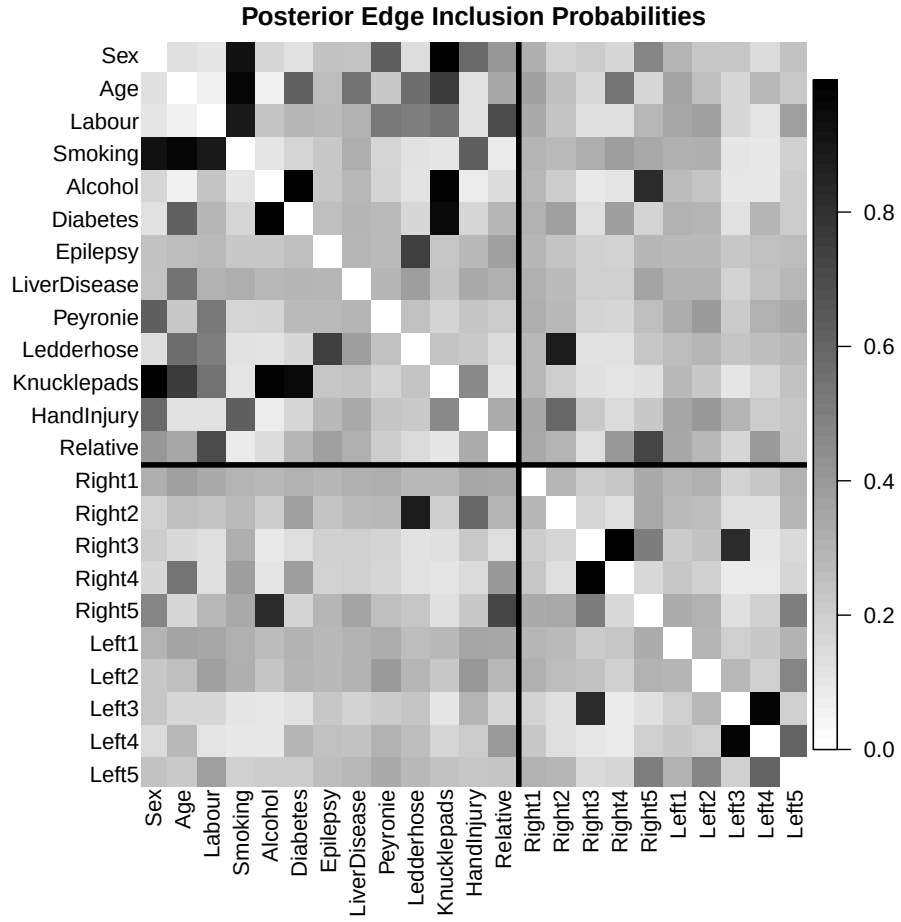


Fig. 3.5 Image visualization of the posterior pairwise edge inclusion probabilities of all possible edges in the graph, for 10 fingers with 14 risk factors.

fingers are almost independent in terms of the severity of Dupuytren disease, there is no reason to operate the fingers simultaneously.

To apply our Bayesian framework for these 10 variables (fingers), we place a uniform distribution as an uninformative prior on the graph and the G-Wishart $W_G(3, I_{10})$ on the precision matrix. We run our BDMCMC algorithm for $2,000$K iteration with a $1,000$K as burn-in.

The graph with the highest posterior probability is the graph with 12 edges. Figure 3.6 shows the selected graph with 8 edges, for which the posterior inclusion probabilities in (4.12) is greater then 0.5. Edges in the graph show the interactions among the variables under consideration. Figure 3.7 shows the table of all posterior inclusion probabilities.

The results (Figures 3.6 and 3.7) show that there are significant co-occurrences of Dupuytren disease in the ring fingers and middle fingers in both hands. Therefore we can infer that middle finger substantially associate to the ulnar side of the hand. Surpris-

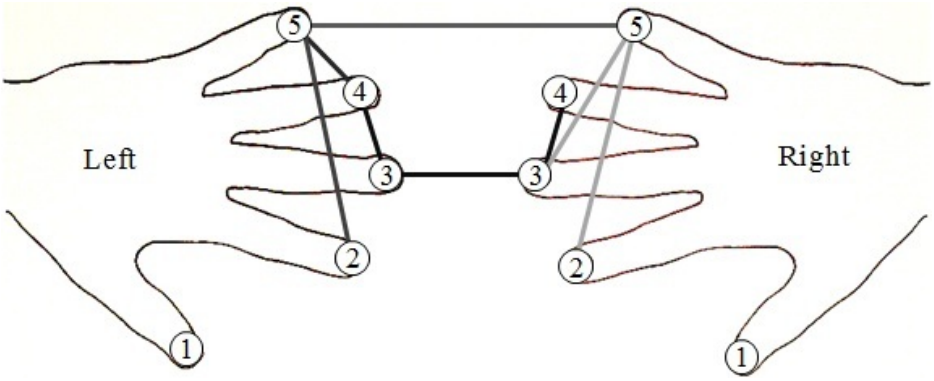


Fig. 3.6 The inferred graph the Dupuytren disease dataset based on the total degrees of flexion in all 10 fingers. It reports the selected graph with 9 significant edges for which their posterior inclusion probabilities (4.12) are more than 0.5.

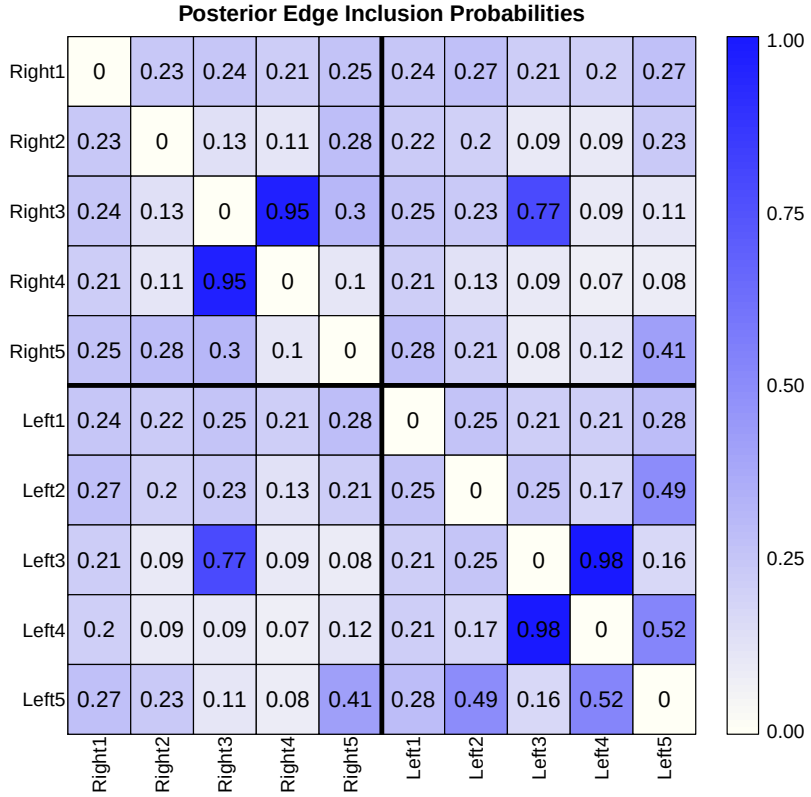


Fig. 3.7 Image visualization of the posterior pairwise edge inclusion probabilities of all possible edges in the graph, for 10 fingers.

ingly, our result shows that there is a significant relationship between middle fingers in both hands. This result supports the hypotheses the the disease has genetic factors or other biological factors that affect fingers simultaneously. Moreover, it also shows that the joint interactions between fingers in both hand is almost symmetric.

3.4.3 Fit of model to Dupuytren data

Posterior predictive checks can be used for checking the proposed Bayesian approach fits the Dupuytren data set. If the model fits the Dupuytren data set, then simulated data generated under the model should look like to the observed data.

Therefore, first, based on our estimated graph from our BDMCMC algorithm in section 3.4.1, we draw simulated values from the posterior predictive distribution of replicated data. Then, we compare the samples to our observed data. Any systematic differences between the simulations and the data determine potential failings of the model.

In this regard and based on the result in subsection 3.4.1 (Figures 3.4 and 3.5), for both simulation and observed data, we obtain the conditional distributions of the potential risk factors and fingers. We show that the result for finger 4 in right hand and risk factor age in figure 3.8, for finger 5 in right hand and risk factor relative in figure 3.9, and for finger 2 in right hand and risk factor Ledderhose in figure 3.10.

Figure 3.8 plots the empirical and predictive distribution of variable finger 4 in hand right conditional on variable "age"in four categories $\{(40, 50), (50, 60), (60, 70), (70, 90)\}$. For variable finger 4 in hand right, based on Tubiana Classification, we categories it in 5 categories (category 1: 0 degree for total angle; 2: degree between $(1, 45)$; 3: degree between $(46, 90)$; 4: degree between $(90, 135)$; 5: degree more than 135).

Figure 3.9 plots the empirical and predictive distribution of variable finger 5 in hand right conditional on variable "Relative".

Figure 3.10 plots the empirical and predictive distribution of variable finger 2 in hand right conditional on variable "Ledderhose". Figures 3.10 and 3.9 show that the fit is good, since the predicted conditional distributions, in general, are the same as the empirical distributions.

3.5 Conclusion

In this paper we have proposed a Bayesian methodology for discovering the effect of potential risk factors of Dupuytren disease and the underling relationships between fingers simultaneously.

The results of the case study clearly demonstrate that age, alcohol, relative, and ledderhose diseases all affect Dupuytren disease directly. Other risk factors only affect Dupuytren disease indirectly. Another important result is that severity of the Dupuytren disease in fingers are correlated. In particular, the middle finger with the ring finger. This implies that

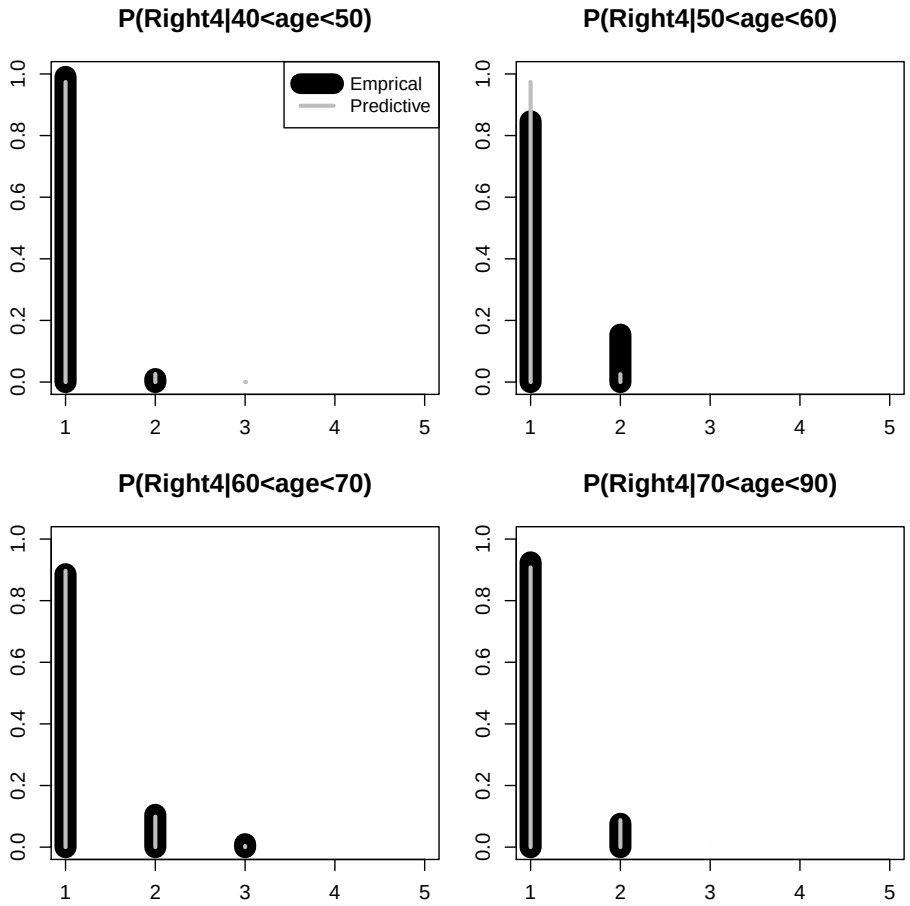


Fig. 3.8 Empirical and predictive conditional distributions for total angle of finger 4 in right hand condition on four different categories of variable "age".

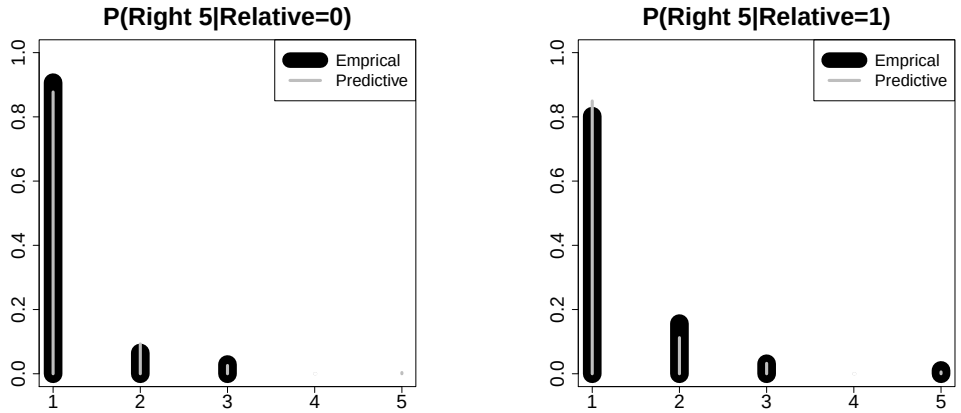


Fig. 3.9 Empirical and predictive conditional distributions for total angles of finger 5 in right hand condition on relative variable.

a surgical intervention on either the ring or middle finger should preferably be executed simultaneously.

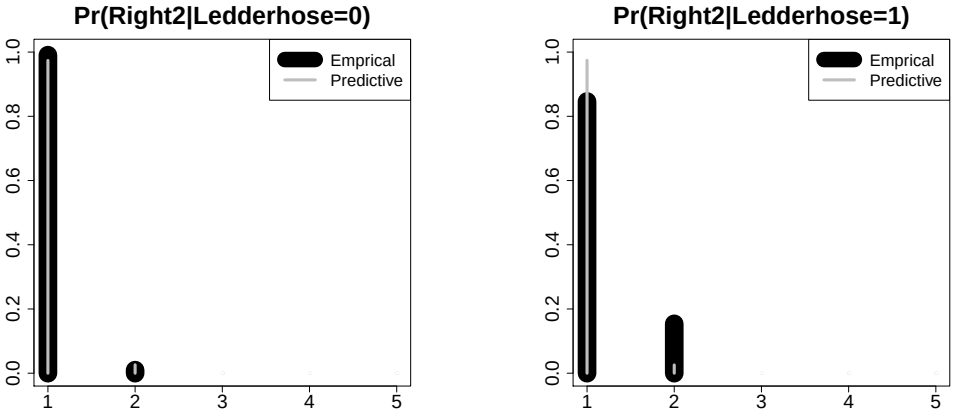


Fig. 3.10 Empirical and predictive conditional distributions for total angles of finger 2 in right hand condition on Ledderhose disease variable.

Of course, our proposed Bayesian methodology is not limited only to this type of data. It can potentially be applied to any kind of mixed data where the observed variables are binary, ordinal or continuous. Our method does not work with discrete variables that are not binary or ordinal.

We compare our Bayesian approach with an alternative Bayesian approach (7) using a simulation study on various types of network structures. Although, both approaches converge to the same posterior distribution our approach has some clear advantages on finite MCMC runs. This difference is mainly due to our implementation of a computationally efficient algorithm. Our method is computationally more efficient because of two reasons. Firstly, our sampling algorithm is based on birth-death process which compare to the RJMCMC implemented in (7) is much more efficient. Secondly, based on the theory which is derived by (32), we used exact values for the ratio of normalizing constants which has been computationally a bottleneck in the Bayesian approach. Moreover, we implemented the code for our method in C++ which is linked to R; it is freely available online by R package `BDgraph` at http://CRAN.R-project.org/package=BDgraph.

References

[1] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

[2] Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335.

[3] Bayat, A. and McGrouther, D. (2006). Management of dupuytren's disease–clear advice for an elusive condition. *Annals of the Royal College of Surgeons of England*, 88(1):3.

[4] Cappé, O., Robert, C., and Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700.

[5] Cheng, Y., Lenkoski, A., et al. (2012). Hierarchical gaussian graphical models: Beyond reversible jump. *Electronic Journal of Statistics*, 6:2309–2331.

[6] Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

[7] Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.

[8] Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433.

[9] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

[10] Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283.

[11] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400.

[12] Lanting, R., Broekstra, D. C., Werker, P. M., and van den Heuvel, E. R. (2014a). A systematic review and meta-analysis on the prevalence of dupuytren disease in the general population of western countries. *Plastic and reconstructive surgery*, 133(3):593–603.

[13] Lanting, R., Nooraee, N., Werker, P., and van den Heuvel, E. (2014b). Patterns of dupuytren disease in fingers; studying correlations with a multivariate ordinal logit model. *Plastic and reconstructive surgery*.

[14] Lanting, R., van den Heuvel, E. R., Westerink, B., and Werker, P. M. (2013). Prevalence of dupuytren disease in the netherlands. *Plastic and reconstructive surgery*, 132(2):394–403.

[15] Lauritzen, S. (1996). *Graphical models*, volume 17. Oxford University Press, USA.

[16] Lenkoski, A. (2013). A direct sampler for g-wishart variates. *Stat*, 2(1):119–128.

[17] Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157.

[18] Liang, F. (2010). A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

[19] Meyerding, H. W. (1936). Dupuytren's contracture. *Archives of Surgery*, 32(2):320–333.

[20] Milner, R. (2003). Dupuytren's disease affecting the thumb and first web of the hand. *Journal of Hand Surgery (British and European Volume)*, 28(1):33–36.

[21] Mohammadi, A. and Wit, E. (2015a). *BDgraph: Graph Estimation Based on Birth-Death MCMC Approach*. R package version 2.17.

[22] Mohammadi, A. and Wit, E. C. (2015b). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

[23] Mohammadi, A. and Wit, E. C. (2015c). Bdgraph: Bayesian structure learning of graphs in r. *arXiv preprint arXiv:1501.05108v2*.

[24] Muirhead, R. (1982). *Aspects of multivariate statistical theory*, volume 42. Wiley Online Library.

[25] Murray, I., Ghahramani, Z., and MacKay, D. (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

[26] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

[27] Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.

[28] Scutari, M. (2013). On the prior and posterior distributions used in graphical modelling. *Bayesian Analysis*, 8(1):1–28.

[29] Shih, B. and Bayat, A. (2010). Scientific understanding and clinical management of dupuytren disease. *Nature Reviews Rheumatology*, 6(12):715–726.

[30] Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

[31] Tubiana, R., Simmons, B., and DeFrenne, H. (1982). Location of dupuytren's disease on the radial aspect of the hand. *Clinical orthopaedics and related research*, (168):222.

[32] Uhler, C., Lenkoski, A., and Richards, D. (2014). Exact formulas for the normalizing constants of wishart distributions for graphical models. *arXiv preprint arXiv:1406.4901*.

[33] Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

[34] Wang, H. (2014). Scaling it up: Stochastic search structure learning in graphical models.

[35] Wang, H. and Li, S. (2012). Efficient gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198.

[36] Wang, H. and Pillai, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *Journal of Computational and Graphical Statistics*, 22(3):689–707.

[37] Wong, F., Carter, C. K., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830.